PCA practical tips

Tip 1 → Make sure the variables are on the same scale and if not scale them.

Suppose let take dataset, Aptitude test

	Student 1	S2	S3	S4	...
Math	95	88	93	75	...
English	9	8	10	7	...

→ Math score is spread from 0 to 100 and English score from 0 to 10.
If we applied PCA, PC1 = 0.99 (Maths) + 0.1 (English)
→ This suggest that maths is 10 times better than English for capturing variation in aptitude test. But this is only because the math score are on a scale 10 times larger than scale of English scores.
→ So solution is scaling, divide the Math score by 10 so that it will be on same scale. So now after scaling PC1 = 0.77 (Math) + 0.77 (English)
This suggests that English & Maths are equally good at capturing variation in aptitude.

- So in short, we need to make sure the scales for each variable (in this case maths and english) are roughly equivalent, otherwise we will be biased toward one of them.
- Standard practise is to divide each variable by its standard deviation. Thus if a variable has wide range, it will have a large standard deviation and dividing by it will scale the value a lot. If a variable has a narrow range, it will have a small standard deviation and scaling will be minimal.

Practical tip 2 → How many principal component should we expect?

So if we have 2 variables, means we can plot 2 variable plot / 2D plot.
- So at max we can have 2 perpendicular line. in 2D graph. we cannot have third perpendicular line (which will be perpendicular to both 2 perpendicular line)
- 3 perpendicular line is possible when we have 3D or above 3D graph. So for atleast 3D we need minimum of 3 variables.
- So for 2 variables we can have maximum of 2 pc's.
- Suppose if we have 2 variables which are 100% correlated. so after we do PCA, PC1 will have 100% variation which proves that we need only one variable instead of two variable which explains 100% variation.
- In summary, technically there is a PC for each variable in the dataset. However if there are fewer samples than variables, then number of sample should be number of PCs.